Table of Contents

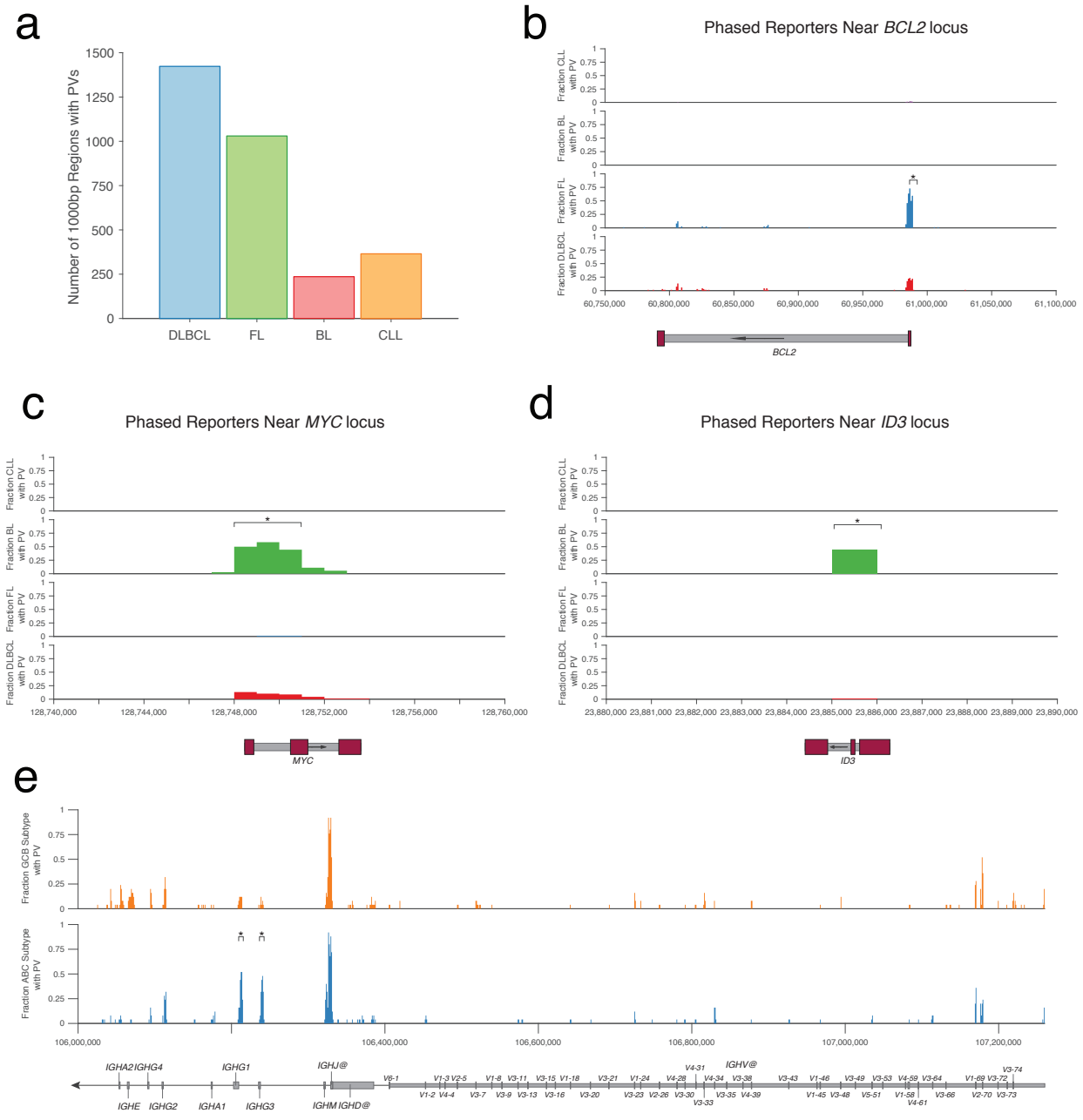

Figure S1. Quantity and genomic location of PVs from WGS in lymphoid malignancies. a) Bar plot showing the number of independent 1000bp regions across the genome that recurrently contain PVs for DLBCL, FL, BL, and CLL (n=68, 74, 36, and 151 respectively). b-d) Plots showing the frequency of PVs for multiple lymphoid malignancies with relationships to specific genetic loci, including b) *BCL2*, c) *MYC*, and d) *ID3*. The location of the transcript for a given gene is shown below the plot in grey; exons are shown in red. * indicates regions with

significantly more PVs in a given cancer histology compared to all other histologies by Fisher's Exact Test ($P<0.05$).

e) Similar to b-d), these plots show the frequency of PVs across lymphoma subtypes. Here, we show the *@IGH* locus, consisting of *IGHV*, *IGHD,* and *IGHJ* parts, for ABC and GCB subtype DLBCLs (n=25 and 25, respectively). Coding regions for Ig parts, including Ig-constant regions and V-genes, are shown. * indicates regions with significant difference in PVs between ABC and GCB subtypes by Fisher's Exact Test ($P<0.05$). (DLBCL, diffuse large B-cell lymphoma; FL, follicular lymphoma; BL, Burkitt lymphoma, CLL, chronic lymphocytic leukemia)

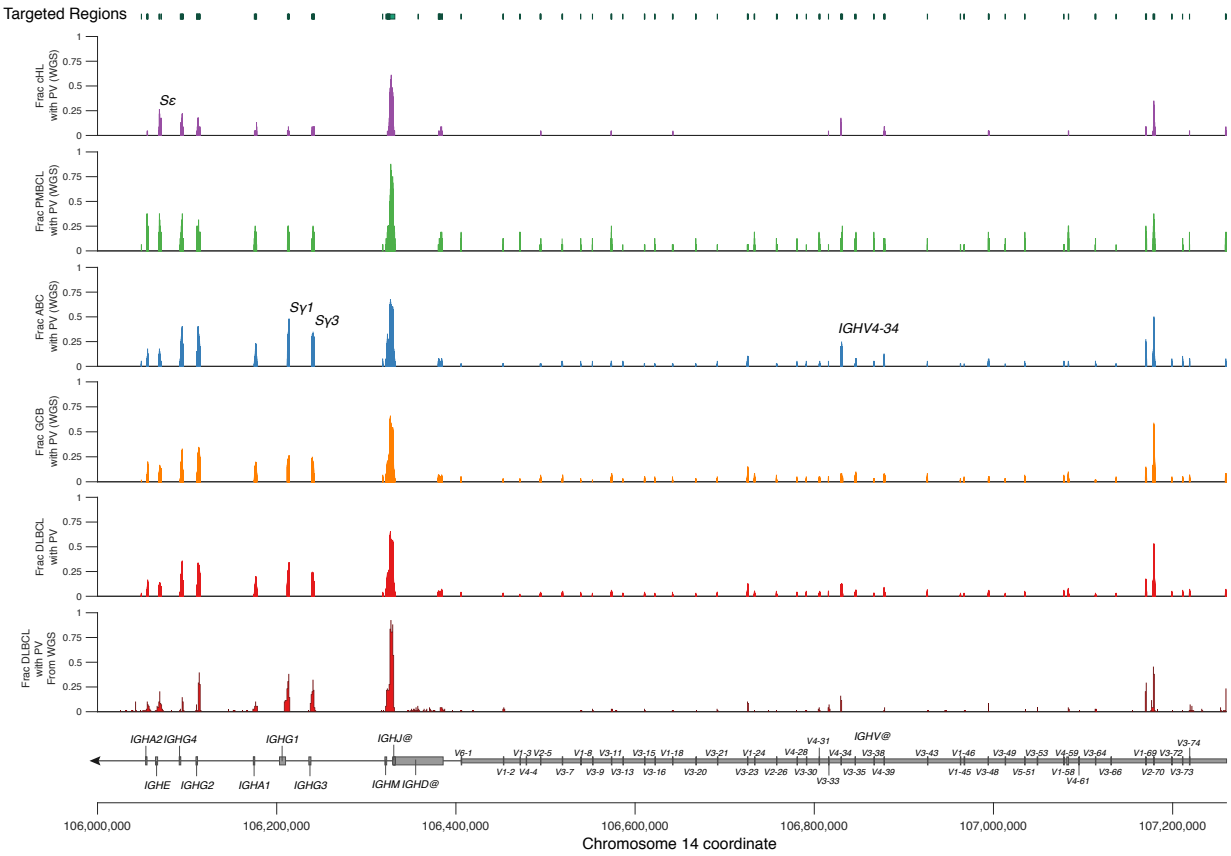

Figure S2. Differences in PVs between lymphoma types in mutations in the *IGH* locus. This figure shows the frequency of PVs from PhasED-Seq across the *@IGH* locus for different types of B-cell lymphomas. The bottom track (grey) shows the structure of the *@IGH* locus and gene-parts, including Ig-constant genes and V-genes. The next (red, outlined) track shows the frequency of PVs in this genomic region from WGS data (ICGC cohort). The remainder of the tracks show the frequency of PVs from PhasED-Seq targeted sequencing data, including 1) DLBCL (red), GCB-DLBCL (orange), ABC-DLBCL (blue), PMBCL (green), and HL (purple). The regions targeted by the PhasED-Seq panel are shown at the top. Selected immunoglobulin parts with PVs enriched in specific histologies are labeled (i.e., *IGHV4-34*, *Sε*, *Sγ3* and *Sγ1*).

Supplementary Tables (Attached in separate Excel file)

Table S1. WGS data for discovery of phased variants.

This table contains the relevant samples and data sources from PCAWG and ICGC used for discovery of putative phased variants (PVs) in WGS from 2538 cases across 24 histologies of cancer. The number of SNVs (numSNVs), putative PVs for both two and three variants in phase (numPVs 2x, numPVs 3x), and enrichment of PVs, defined as number of doublet PVs / number of SNVs (pvEnrichment) are also shown.

Table S2. Regions of recurrent PVs in lymphoid malignancies.

This table shows the frequency of putative PVs in 1-kb regions across the genome in lymphoid malignancies. Only regions with >1 subject containing a putative PV are shown.

Table S3. Technical comparison between CAPP-Seq and PhasED-Seq.

This table shows the sequencing depth, SNV counts, and PV counts for samples sequenced by both CAPP-Seq and PhasED-Seq panels.

Table S4. Clinical characteristics of cases used for PV enrichment across PhasED-Seq panel.

Table S5. Enrichment of PVs across regions of PhasED-Seq panel according to lymphoma subtype.

This table shows the enrichment in PVs at specific genetic loci for lymphoma subtypes throughout the PhasED-Seq panel. Genetic loci were determined by separating the panel into 50bp bins and assigning each bin to a specific gene based on the nearest GENCODE gene annotation.

Table S6. Oligonucleotides used for hybridization bias experiment.

This table provides the sequences of oligonucleotides synthesized to assess hybridization and molecular recovery bias with increasing mutational burden.

Table S7. Technical assessment of PhasED-Seq limit of detection through limiting dilution series.

This table provides the sequencing characteristics including deduplicated read count and depth for samples used in the limiting dilution experiment establishing the limit of detection for PhasED-Seq. Three separate patient cfDNA samples were spiked into healthy control cfDNA and serially diluted to expected tumor content levels of 1 in 1,000 to 1 in 2,000,000. The allele fraction of each sample assessed by standard CAPP-Seq, duplex sequencing, and PhasED-Seq using doublets and triplets, is also provided. Allele fractions were normalized to the tumor fraction of the highest sample concentration (i.e., the 1:1,000 sample). Background detection of patient-specific PV reporters in control samples from 12 standard depth and 1 deep control sample are also shown.

Table S8. Clinical characteristics of cases used for evaluation of MRD performance by lymphoma PhasED-Seq.

This table provides the clinical characteristics of lymphoma patient samples used for assessment of MRD detection at cycle 2 day 1, cycle 3 day 1, and/or end of therapy.

Table S9. Characteristics of cases used for personalized PhasED-Seq in solid tumors.

Clinical characteristics of the cases used for extension of PhasED-Seq to solid tumors via personalized approach. The stage, relevant treatment, and plasma samples utilized are also shown. All samples were assessed for ctDNA by SNV-based CAPP-Seq for comparison as well.

Supplementary Methods

Identification of phased variants

Identification of phased variants and allelic quantitation

After generating UMID error-suppressed alignment files (i.e., BAM files), we identified PVs from each sample as follows. First, matched germ-line sequencing of uninvolved peripheral blood mononuclear cells (PBMCs) was performed to identify patient-specific constitutional single nucleotide polymorphisms (SNPs). These were defined as non-reference positions with a variant allele fraction (VAF) above 40% with a depth of at least 10, or a VAF of above 0.25% with a depth of at least 100. Next, we identified PVs from read-level data for a sample of interest. Following UMID-mediated error suppression, we analyzed each individual paired-end (PE) read and identified all non-reference positions using 'samtools calmd'. We used PE data rather than single reads to identify variants occurring on the same template DNA molecule, which could subsequently fall into either read 1 or read 2. Any read-pair containing ≥2 non-reference positions was considered to represent a possible somatic PV. For reads with >2 non-reference positions, each permutation of size ≥2 was considered independently: i.e., if 4 non-reference positions were identified in a read-pair, all combinations of 2 SNVs (i.e., 'doublet' phased variants) and all combinations of 3 SNVs (i.e., 'triplet' phased variants) were independently considered. PVs containing putative germ-line SNPs were also removed as follows: if in a given *n*-mer (i.e., *n* SNVs in phase on a given molecule) ≥*n*-1 of the component variants were identified as germ-line SNPs, the PV was redacted. This filtering strategy ensures that for any remaining PV, at least 2 of the component SNVs were not seen in the germ-line, as relevant for both sensitivity and specificity.

We further filtered putative somatic PVs using a heuristic blacklisting approach in considering sequencing data from 170 germ-line DNA samples serving as controls. In each of these samples, we identified PVs on read-pairs as described above, but without filtering for matched germ-line. Any PV that occurred in one or greater paired-end read, in one or more of these control samples, was included in the blacklist and removed from patient-specific somatic PV lists.

To calculate the VAF of each PV, we calculated a numerator representing the number of DNA molecules containing a PV of interest over a denominator representing the total number of DNA molecules that covered the genomic region of interest. That is, the numerator is simply the total

number of deduplicated read-pairs that contain a given PV while the denominator is the number of read-pairs that span the genomic locus of a given PV.

Genotyping phased variants from pretreatment samples

The above strategy resulted in a list of PVs of ≥1 read-depth in each sample. To identify PVs serving as tumor-specific somatic reporters for disease monitoring, for each case we identified a 'best genotyping' specimen – either DNA from a tumor tissue biopsy (preferred), or pretreatment cell-free DNA. After identifying all possible PVs in the 'best genotyping sample', we further filtered the list for specificity as follows. For any *n*-mer PV set, if ≥*n*-1 of the constituent SNVs were present as germ-line SNPs in the 170 control samples described above, the PV was removed. Furthermore, only PVs that meet the following criteria were considered: 1) AF > 1%; 2) depth of the PV locus of ≥100 read-pairs, 3) at least one component SNV in the on-target space, 4) zero read-support in matched germline. Finally, 5) any PV meeting these criteria was assessed for read-support in a cohort of 12 healthy control cfDNA samples. If any read-support was present in >1 of these 12 samples, the PV was removed. For genotyping from cell-free DNA samples identified as low tumor fraction by SNVs (i.e., < 1% mean AF across all SNVs), the AF threshold for determining PVs was relaxed to >0.2%. This filtering resulted in the PV lists used for disease monitoring and MRD detection.

Determination of tumor fraction from phased variants

When evaluating a sample for minimal residual disease (MRD) detection with prior knowledge of the tumor genotype, we assessed for the presence of any PV identified in the best pretreatment genotyping sample in the MRD sample of interest. Given a list of *k* possible tumor-derived PVs observed in the best genotyping sample, we first determined all read-pairs covering at least 1 of the *k* possible PVs. This value, *d*, can be thought of as the aggregated 'informative depth' across all PVs spanned by cfDNA molecules in a PhasED-Seq experiment[1]. We then assessed how many of these *d* read-pairs actually contained 1 or more of the *k* possible PVs – this value, *x*, represents the number of tumor-derived molecules containing somatic PVs in a given sample. The number of tumor-derived molecules containing PVs divided by the informative depth – *x*/*d* – is therefore the phased-variant tumor fraction (PVAF) in a given sample. For detection of MRD in each sample, we calculated PVAF independently for doublet and triplet PVs.

Monte Carlo simulation for significance of PV detection

To assess the statistical significance of the detection of tumor-derived PVs in any sample, we implemented an empiric significance testing approach. We first define a test statistic *f* as follows – from a given list of *k* possible tumor-derived PVs observed in the best genotyping sample, we calculate the arithmetic mean of allele fractions across all *k* PVs (allele fraction defined as the number of read-pairs containing an individual PV (x_i) over the number of read-pairs spanning the PV positions (d_i)):

$$\mathrm{f} = \frac{\sum_{i=1}^{k} \frac{x_i}{d_i}}{k} \quad (1)$$

to assess the hypothesis that *f* is not significantly different from the background error-rate of similar PVs assessed from the same sample. We used a Monte Carlo approach to develop a null distribution and perform statistical testing as follows:

1. Given a set of *k* PVs, $\{pv_1 \ldots pv_i \ldots pv_k\}$, we generated an 'alternate' list of PVs, $\{pv'_1 \ldots pv'_i \ldots pv'_k\}$, such that for each alternate PV had the same type of base change and distance between SNVs as the test PV. For example, if a doublet PV, chr14:106329929 C>T and chr14:106329977 G>A, was identified in the genotyping sample, we searched for an alternate two positions at the same genomic distance (here, 48bp) with reference bases C and G, and assessed for read-pairs with the same types of base changes (i.e., C>T and G>A), using the heuristic search scheme below.
2. For each tumor pv_i in our set of *k*, we identified 50 such alternates. This was performed with a random search algorithm to scan the targeted space and identify alternates. To find these 50 alternates, we began by identifying a random position on the same chromosome as the test pv_i and then searched for the same types of reference bases at the same genomic distance as described above. Synteny of observed/alternate PVs was used to control for regional variation in SHM/aSHM as well as copy number variation, as potential confounders of the null distribution. Alternate positions that were identified as a germ-line SNP, defined as having AF > 5%, were excluded.
3. After identifying 50 such alternates for each pv_i, we generated 10,000 random permutations of 1 alternate for each of the *k* original PVs and calculated the phased-

variant fraction *f'* for these alternate lists in the sample of interest being evaluated for presence of MRD, as described above.

4. We then calculated an empiric P-value, defined as the fraction of times the true phased-variant fraction *f* is observed to be less than or equal to the alternate *f'* across the 10,000 random PV lists as an empirical measure of significance of MRD significance in the blood sample of interest.
5. This empiric P-value is calculated independently for both doublet and triplet PVs (i.e., two and three SNVs in phase). These P-values are then combined using Fisher's method to result in a composite signal for tumor content.

While this resulting comparison is a measure of the significance for PV detection of our tumor-reporter list compared to the empirically defined background PV error-rate *within* the sample of interest, we also evaluated its relationship to specificity of detection *across* cases and control samples, as described below.

Assessment of specificity of PhasED-Seq

To determine the specificity of disease and MRD detection through PhasED-Seq, we first identified patient-specific PVs from 107 patients with DLBCL using pretreatment tumor or plasma DNA along with paired germ-line samples. We then assessed 40 independent plasma DNA samples from healthy individuals for presence of these patient-specific PVs, using the Monte Carlo approach outlined above. We empirically determined a P-value threshold from our Monte Carlo to ensure high specificity; in the lymphoma cohort, this resulted in a P-value threshold of 0.01 yielding 97% specificity, with 98% sensitivity in pretreatment samples (**Fig Extended Data Fig 6a**).

Calculation of error rates

To assess the error profile of both isolated SNVs and PVs, we examined the non-reference base observation rate of each type of variant across all reads. For isolated SNVs, we calculated the error-rate for each possible base change $e_{n1>n1'}$ as the fraction of on-target bases with reference allele $n1$ that are mutated to alternate allele $n1'$, when considering all possible base-changes of the reference allele. Positions with a non-reference allele rate exceeding 5% were classified as probable germ-line events, and excluded from the error-rate analysis. We also

calculated a global error rate, defined as the rate of mutation from the hg19 reference allele to any alternate allele.

For phased variants, we performed a similar calculation. For the error-rate of a given type of phased variant composed of *k* constituent base-changes $\{e_{n1>n1'} \ldots e_{nk>nk'}\}$, we calculated the error-rate by determining both the number of instances of the type of base change (i.e., the numerator), as well as the number of possible instances for the base change (i.e., the denominator). To calculate the numerator, *N*, we counted the number of occurrences of the PV of interest over all read-pairs in a given sample. For example, to calculate the error-rate of C>T and G>A phased doublets, we first counted the number of read-pairs that include *both* a reference C mutated to a T as well as a reference G mutated to an A.

To calculate the denominator, *D*, we also calculated the number of possible instances of this type of phased variant; this was performed first for each read-pair *i*, and then summed over all read pairs. A PV with *k* components can be summarized as having certain set of reference bases p_A, p_C, p_G, p_T, where p_N is the number of each reference base in the PV. Similarly, a given read pair contains a certain set of reference bases b_A, b_C, b_G, b_T, where b_N is the number of each reference base in the read pair. Therefore, for each read pair in a given sample, the number of possible occurrences of our PV type of interest can be calculated combinatorically as:

$$D_i = \binom{b_A}{p_A}\binom{b_C}{p_C}\binom{b_G}{p_G}\binom{b_T}{p_T} \quad (2)$$

For example, consider a read-pair with 40 reference As, 50 reference Cs, 45 reference Gs, and 35 reference Ts. The number of positions for a C>T and G>A PV is:

$$D_i = \binom{40}{0}\binom{50}{1}\binom{45}{1}\binom{35}{0} = 2250 \quad (3)$$

The aggregated denominator, *D*, for error rate calculation is then simply the sum of this value over all read pairs. The error rate for this type of PV is then simply *N/D*.

Differences in phased variants between lymphoma subtypes

To compare the distribution of phased variants in different types of lymphomas, we identified tumor-specific PVs in 101 DLBCL, 16 PMBCL, 23 cHL, 13 FL, and 13 MCL patients via sequencing of tumor biopsy specimens and/or pre-treatment cell-free DNA and paired germ-line specimens. The clinical characteristics of these patients are detailed in **Table S4.** After identifying these tumor-specific PVs, we then assessed their distribution across the targeted sequencing panel. We began by dividing our panel into 50bp bins; for each patient, we then determined if each patient had evidence of a PV within the 50bp bin, defined as having at least one component of the PV within the bin. We furthermore determined the nearest gene to each 50bp bin, based on GENCODEv19 annotation of the reference genome.

To assess how the distribution of PVs between subtypes of lymphoma varies at the level of specific genes, we examined the distribution of PVs across the 50bp bins spanning each gene (or nearest gene). For example, consider a given gene with *n* such 50bp bins represented in our targeted sequencing panel. For each bin, we first determined the fraction of patients, *f*, in each type of lymphoma with a PV falling within the 50bp bin – i.e., we determine $\{f_{type1,1}, \ldots f_{type1,n}\}$ and $\{f_{type2,1}, \ldots f_{type2,n}\}$. We then compared any two histologies for the fraction of cases harboring PVs in the set of 50bp bins assigned to each gene. The number of 50bp bins in each gene as well as summary statistics for each histology are provided in **Table S5**. These comparisons are depicted for individual representative genes in **Fig 3c**, with additional figures available for all significantly differentially effected genes available at https://phasedseq.stanford.edu (description of significance testing outlined below).

We statistically compared the enrichment in PVs in a specific lymphoma type or subtype vs. another by calculating the difference in the fraction of patients which contain a PV in each 50bp bin across all bins assigned to a gene (i.e., overlapping a given gene or with a given nearest gene). Specifically, for any comparison between two lymphoma types ($type_1$ and $type_2$), we first identified this set of differences in PV-rate between histologies $\{f_{type1,1} - f_{type2,1}, \ldots f_{type1,n} - f_{type2,n}\}$. We then compared this set of gene-specific differences in frequency of PVs between types of lymphoma against the distribution of all other 50bp bins in the sequencing panel by the Wilcoxon rank sum test. For this test, we compared the set of *n* 50bp bins assigned to a given gene to all other 50bp bins (i.e., 6755 – *n*, since there are 6755 50bp bins in our sequencing panel). This P-value, along with the mean difference in fraction of patients with a PV in each bin

for each gene between histologies, is depicted as a volcano plot in **Fig 3d**. To account for the global difference in rate of PVs between different histologies, the mean difference in fraction of patients with a PV between histologies was centered on 0 by subtracting the mean difference across all genes.

Model to predict the probability of detection for a given set of phased variants

To build a mathematical model to predict the probability of detection for a given sample of interest, we began with the common assumption that cfDNA detection can be considered a random process based on binomial sampling[2]. However, unlike SNVs occurring at large genomic distances apart from one another, detection of PVs can be highly inter-dependent, especially when PVs are degenerate (i.e., when two PVs share component SNVs) or occur in close proximity. To account for this, we only considered PVs occurring >150bp apart from each other as independent 'tumor reporters' (**Extended Data Fig 4f**). The number of 'tumor reporters' to allow for disease detection in a given sample can thus be determined as follows. The PhasED-Seq panel was broken apart into 150bp bins. Each PV in a given patient's reporter list was then turned into a BED coordinate, consisting of the start position (defined as the left-most component SNV) and end position (defined as the right-most component SNV). For each PV, the 150bp bin from the PhasED-Seq selector panel containing the PV was determined; if a PV spanned two or more 150bp bins, it was assigned to both bins. The number of independent tumor reporters was then defined as the number of separate 150bp bins containing a tumor-specific PV.

We then developed a mathematical model comparing the expected probability of detection for a given sample at a given tumor fraction with a given number of independent tumor reporters (i.e., 150bp bins). With a given number of tumor reporters r, at a given tumor fraction f, with a given sequencing depth d, the probability of detecting 1 or more cell-free DNA molecule containing a tumor-specific PV containing can be defined as:

$$Pr(detection) = 1 - Pr(nondetection) \quad (4)$$

$$= 1 - \binom{d * r}{0} f^0 (1-f)^{d*r} \quad (5)$$

based on simple binomial sampling. However, as we trained our ctDNA detection method to have a 5% false positive rate, we add this false positive rate term to our model as well:

$$Pr(detection) = 1 - Pr(nondetection) + 0.05 * Pr(nondetection) \quad (6)$$

$$Pr(detection) = 1 - 0.95 * \ Pr(nondetection) \quad (7)$$

$$= 1 - 0.95 * \binom{d * r}{0} f^0 (1-f)^{d*r} \quad (8)$$

Extended Data Fig 5h shows the results of this model for a range of tumor reporters r from 3 to 67 at depth d of 5000. The confidence envelope on this plot shows solutions for a range of depth d from 4000 to 6000.

To empirically validate this model assessing the probability of disease detection, we utilized samples from our limiting dilution series. In this dilution series, 3 patient cfDNA samples, each containing patient-specific PVs, were spiked into healthy control cfDNA. For each list of patient specific PVs, we performed 25 random subsamplings of the 150bp bins containing patient-specific PVs to generate reporter lists containing variable numbers of tumor-specific reporters. We selected a maximum bin number of 67 to allow sampling from all 3 patient-specific PV lists, followed by scaling down the number of bins by 2x or 3x per step. This resulted in reporter lists containing patient-specific PVs from 3, 6, 17, 34, or 67 independent 150bp bins. We then assessed for disease detection using each of these patient-specific PV lists of increasing size in each of our 'wet' limiting dilution samples from 1:1,000 to 1:1,000,000 (**Extended Data Fig 5i**, closed circles). We furthermore created *in silico* mixtures using sequencing reads from our limiting dilution samples with varying expected tumor-content, and again assessed for the probability of disease detection using patient-specific subsampled PV reporter lists of varying lengths (open circles). For this experiment, we down-sampled both the 'wet' and '*in-silico*' dilution bam files to achieve a depth of ~4000-6000x to correspond with our modeled depth. The final mean and standard deviation of depth across all down-sampled bam files was 4214x ± 789. The probability of detection was summarized across all tests at a given expected tumor fraction, for a given patient-specific PV list. For each given dilution, we considered multiple independently sampled sets of reads to allow superior estimation of the true probability of

detection. Specifically, we considered the following number of replicates at each dilution indicated:

Dilution	**Replicates**	**Number of Tests (Replicates * 25)**	**Wet or *In silico***
1 : 1,000	1	25	Wet
5 : 10,000	3	75	*In silico*
3.5 : 10,000	3	75	*In silico*
2 : 10,000	3	75	*In silico*
1 : 10,000	3	75	Wet
5 : 100,000	3	75	*In silico*
3.5 : 100,000	3	75	*In silico*
2 : 100,000	3	75	*In silico*
1 : 100,000	3	75	Wet
5 : 1,000,000	8	200	*In silico*
3.5 : 1,000,000	8	200	*In silico*
2 : 1,000,000	8	200	Wet
1 : 1,000,000	8	200	Wet

The total number of tests, for each patient-specific PV list, is therefore the number of randomly subsampled PV lists (e.g., 25) times the number of independently downsampled bam files; this number is provided in the table above. In **Extended Data Fig 5i**, the points and error-bars represent the mean, minimum, and maximum across all three patients. The concordance between the predicted probability of disease detection from our theoretical mathematical model and our wet and *in silico* samples validating this model, is shown in **Extended Data Fig 5j**.

Model to assess theoretical sensitivity of ctDNA by tracking PVs vs other variants

To assess the utility of tracking phased variants for ctDNA in diverse cancers, we assess the theoretical sensitivity that would be afforded by developing a personalized sequencing panel for each patient to track a specific set of variants. To do this, we need to know the expected size of a personalized sequencing panel, which depends on the number of variants targeted, as well as the expected depth after UMID-mediated deduplication for a given number of sequencing reads.

We utilized the data from a recently published optimized hybrid capture workflow[3] to infer the expected deduplicated or duplex depth from a given amount of sequencing, for a given panel size, from a given amount of starting cfDNA. Specifically, the above optimized workflow describes a given amount of depth (UMID deduped or duplex) per ng of DNA input. We fit an exponential-asymptotic model to this data, of the form:

$$depth = N * \left(1 - e^{-k\frac{r}{1{,}000{,}000}}\right) \quad (9)$$

Where *N* is the carrying capacity (i.e., the maximum depth that can be recovered) in depth/ng DNA input, and *r* is the number of reads given to a sample per ng DNA input per bp of selector. The result of this model-fitting is shown in **Extended Data Fig 10a**, and resulted in the following coefficients:

	N (maximum depth; x / ng DNA input)	**k (exponential coefficient)**
Total De-duplicated Depth	212.7	2.7e5
Duplex Depth	102.8	4.7e4

Using this model, we then assessed the theoretical sensitivity of possible personalized cfDNA assays considering capturing: 1) phased variants, 2) SNVs utilizing duplex sequencing, or 3) structural variants. In each case from PCAWG, we considered a hypothetical panel covering the maximum number of the variant type of interest (up to a maximum of 10,000), requiring a hybrid capture tile of 120bp (e.g., a panel tracking 100 variants would be 12,000bp, etc). We then assessed the theoretical number of recovered evaluable fragments from this panel using 64ng of DNA input and 20,000,000 sequencing reads, as inferred by the above model. For each case from the PCAWG set, we then assessed which methodology would result in the greatest number of recovered cfDNA fragments evaluable for tumor content[1]. The inferred detection-limit was then considered to be the lowest allele fraction predicted to be detectable with 95% probability based on binomial sampling. The inferred detection-limit for personalized PhasED-

Seq is shown in **Fig 7a**, while the comparison to duplex sequencing or tracking structural variants is shown in **Extended Data Fig S10b-c**.

References

1 Wan, J. C. M. *et al.* ctDNA monitoring using patient-specific sequencing and integration of variant reads. *Science translational medicine* **12**, doi:10.1126/scitranslmed.aaz8084 (2020).

2 Newman, A. M. *et al.* An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nature medicine* **20**, 548-554, doi:10.1038/nm.3519 (2014).

3 Chabon, J. J. *et al.* Integrating genomic features for non-invasive early lung cancer detection. *Nature* **580**, 245-251, doi:10.1038/s41586-020-2140-0 (2020).